Extraction and Quantification of Pack-years and Classification of Smoker Information in Semi-structured Medical Records

Abstract

Electronic medical records contain a wealth of information that is potentially invaluable to many interested parties. However, the fact that most of these documents are of semi-structured nature and are comprised of fragmented English free text, region-specific templates and clinical sublanguage among many other things, has made it difficult to use existing Natural Language Processing tools on them directly and to extract those information. In this work, we focus our attention on a set of medical records pertaining to Rheumatoid Arthritis patients and we present a pattern-based methodology for extracting and quantifying pack-year information. We also introduce an extension to those patterns in classifying individual instances within these documents into a set of predefined smoker status classes. Since our effort in extracting pack-years from medical documents is the first in its kind to the best of our knowledge, we evaluate our approach on a manually selected document collection and present very promising results. We also evaluate our instance classification approach using an additional document collection and show that the approach genaralizes well in both these collections.

1. Introduction

Electronic medical records (EMRs) contain a wealth of data about a variety of aspects including diagnoses, medications and procedures. Consequently, EMRs have been acknowledged as a potential source for identifying a large number of avenues for research studies across many disciplines. However, mining these documents manually for information is quite a tedious task and thus opens up many interesting research paths in the area of Natural Language Processing (NLP). A significant effort can be saved if NLP tools can be applied with minimal supervision on these documents for tasks such as information extraction, quantification and classification. Understanding and reasoning about the structure of these medical documents therefore becomes paramount to such tasks.

However, applying NLP tools directly on these semi-structured EMRs is a quite difficult task since the nature of these records are significantly different from that of many other document sets we encounter in other NLP tasks, such as news articles. These records, primarily written by physicians, mainly consist of fragmented English free text and have semi-structured formats. In addition to that, these documents contain a significant amount of region-specific templates (such as option lists, checkboxes, etc), clinical sublanguage, loose grammer and loose punctuation, spelling mistakes and idiosyncratic acronyms that people use freely without abiding by any standards. The inclusion of these aspects makes it difficult for the existing tools to be directly applied on them.

NLP tools have previously been used to extract data from EMRs for the identification of subjects with diseases including asthma, diabetes mellitus and heart failure. However, smoking history information has been cited as the most challenging to extract from EMRs (Zeng et al., 2006; Zhou et al., 2005), mainly due to their differences from other well-formed free text that were mentioned above. Although there have been many efforts in identifying smoker status (i.e. whether a smoker is a current, past or a non-smoker), little or no research has been done in terms of extracting and quantifying other smoking-related information from EMRs using NLP tools.

1.1. Motivation

Tobacco-smoking has been proved to have an adverse effect on patients with early Rheumatoid Arthritis (RA). It has been linked to the onset of RA and the effects of smoking on RA have been well documented. Thus, our motivation for this research rose from two main directions. Being able to study the

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

relationship between tobacco-smoking patterns and health aspects of RA patients may be very beneficial to the medical community and researchers interested in studying these patterns. If quantitative information can be extracted from these abundance of EMR data, that could provide many insights to the relationship between tobacco-smoking and related aspects of patients' health such as diagnosis, medication and even prognosis. Previously, researchers have tried to capture the relationship between pack-years (a quantification of a patient's tobacco-smoking history) and a patient's health for varying reasons of interest. (Langhammer et al., 2000) conducted a study where they showed that women are more likely to have respiratory problems than men given that both have the same number of pack-years. Several other studies (Trevathan et al., 1983; Valdes et al., 2005; Kadunce et al., 1991; Kasiske & Klinger, 2000) show that pack-years are linked to many health issues such as Dysplasia, Obesity and Premature Facial Wrinkling.

On the other hand, classifying and quantifying tobacco-smoking information from EMRs pose several problems that would be of interest to the NLP community. As mentioned above, the fact that these documents differ from the widely studied NLP document collections makes it interesting to evaluate how well the standard information extraction and classification tools will generalize into such an esoteric document collection. Also the fact that there has been no research efforts previously to extract and quantify tobacco-smoking information from EMRs, makes this an absorbing problem where a subsequent line of research problems could also open up.

1.2. Overview of Proposed Approach

One interesting distinction in this research from other information extraction tasks is that we carry out our research at an instance-level rather than at a document-level. We define an instance to be a character window surrounding a smoking-related term for a predefined number of characters. These smoking-related terms (s-r terms) are defined in a latter section. We chose to select instances rather than complete documents is owing to the fact that we observed different interested parties use isolated information within a document in different ways to make decisions at the document-level. i.e. Given the same information about each of these instances, different groups will use that information differently to arrive at document level decisions. Owing to this reason, we found it most suitable to provide as much information as we possibly could on each instance and leave the ultimate decisions (for example, classifying a document into current, past or non-smoker by evaluating each instance level label within that document) to the relevant groups.

In this study, we focus on extracting and quantifying pack-years from each instance and classifying each instance into a set of predifined classes. A pack-year is generally defined to be the number of cigarette packs a person smokes a day (where a standard pack contains 20 cigarettes) times the number of years the patient smoked. To provide an example, if a patient has smoked 3 packs per day for 40 years he/she has a pack-year value of 120 whereas if a person has smoked 5 cigarettes a day for 10 years then his/her pack-year value would be 2.5 $((5/20) \times 10)$. Since we are not just interested in extracting pack-year phrases from these documents but rather quantifying them and producing an actual value, we approached the problem as one where we can use manually created patterns to extract phrases and subsequently extract and calculate numeric values from those retrieved phrases. We also define patterns to extract year-quit phrases (mentions of the year a patient quit smoking) and use those phrases in later models for classification.

Our classification models are based on word-based-pattern features as well as structure-based-patterns which in combination provides us a rich feature representation for these instances. We utilize Support Vector Machines (SVM), which is the present de-facto standard in the Machine Learning (ML) community for classification. The features are combined with the aforementioned phrases extracted through patterns, to produce a methodology for classifying each instance to a set of predifined classes. For implementations, we make use of Apache's Unstructured Information Management Architecture (UIMA) (Ferrucci & Lally, 2004). UIMA[1] enables us to form a pipeline for these various annotations and provides a well formed modular framework for implementation of such NLP applications.

2. Description of Datasets

For this research, we made use of two document collections. We have used the tobacco smokers document collection that was released as part of the i2b2 NLP challenge in 2006 for identifying smoker status from patient discharge summaries. The document collection (i2b2 document collection hereafter) contains 502 documents. While these documents did not contain perfectly formed sentences, a large portion of these documents' text was grammatically accurate.

As our second document collection, we collected 750

[1] http://uima.apache.org

documents from health administrative stations of the United States Department of Veterens Affairs (VA) belonging to various regions. The documents were contrastingly different to those of i2b2 document collection in that these documents contained a large variety of template-based structures in them. The structures were different from region to region and, for the most part, contained narrative free text. The documents were a mixture of narrative free text and information contained in semi-structured templates. These structures included check boxes, selections and other data types where the patients information could be stored in a semi-structured manner. In addition to that, these documents contain a lot of noise in them in the form of grammatical and punctuation errors such as mis-spellings, unbalanced brackets, etc. Another major issue in such document collections is that most of the deliniating and boundary information is lost when transcribing these documents from their original sources into EMRs. This makes it quite difficult to perform even the basic pre-processing tasks such as tokenizing on these documents.

Understandably, the second document collection (VA documents) poses much more challenges over a well formed document collection. One major challenge is that we are unable to predict the structure of a new document that we might encounter in the future, even if it comes from the same VA domain. This unreliable structure is what makes the information extraction task rather difficult in this domain. As we would present in later sections, extracting information or even classifying these documents are sometimes challenging to a human being. It is in this setting that we make an effort to evaluate how well NLP tools can be adopted to achieve the same task.

3. Previous Approaches

Some of the previous efforts in this area have been due to the increased interest during the i2b2 challenge in identifying smoker information from clinical records in 2006 (Uzuner et al., 2008). The organizers of this competition released a set of 502 de-identified patient records and had annotated the explicit mentions of smoking-related information within the document. The challenge focused on classifying a given clinical document into five classes. The five classes were *Current Smoker*, *Past Smoker*, *Non-Smoker*, *Smoker* and *Unknown*. Following are the definitions provided by the organizers for each of these classes.

Past Smoker A patient whose discharge summary asserts explicitly that the patient was a smoker one year or more ago but who has not smoked for at least one year. The assertion "past smoker" without any temporal qualifications means Past Smoker unless there is text that says that the patient stopped smoking less than one year ago

Current Smoker A patient whose discharge summary asserts explicitly that the patient was a smoker within the past year. The assertion "current smoker" without any temporal qualifications means Current Smoker unless there is text that says that the patient stopped smoking more than a year ago

Smoker A patient who is either a Current or a Past Smoker but whose medical record does not provide enough information to classify the patient as either

Non-Smoker A patient whose discharge summary indicates that he/she never smoked

Unknown A patient whose discharge summary does not mention anything about smoking. Indecision between Current Smoker and Past Smoker does not belong to this category

Several submissions to this competition produced promising results in classifying clinical documents into these five classes. It is worth noting that these submissions only focused on classifying the individual documents into the aforementioned five classes and did not pursue the task of extracting the pack-year values from the documents.

(Aramaki et al., 2006) produced a system where they first extracted smoking-related sentences through a keyword search and applied Okapi-BM25 and k-nearest neighbor clustering. The extracted sentences were judged to be belonging to a class that was most similar to the class of the training sentences using a close similarity measure. A similar phrase level classification using Support Vector Machines (SVMs) was used by (Clark et al., 2008) where they augmented the training set using 1200 additional records corresponding to the same smoker categories. Their results were based on unigram and bigram features which they used for the classifiers. Another (Cohen, 2008) focused on word level features and again used linear SVMs for their classification and added further postprocessing rules to the system.

(Pedersen, 2006) experimented with unsupervised methods where he used Latent Semantic Analysis (LSA) and SenseClusters' second-order representation. In his method, he used bigrams which began with the string 'smok' as features and reported that the unsu-

pervised methods did not perform as well as the supervised methods. (Savova et al., 2008) implemented a hierarchical approach and recast the problem as a sentence classification task. They first filtered out the *unknown* category and worked with a subcorpus of sentences that contained only smoking-related sentences. In the second stage, the *non-smoker* category was identified using a modified version of *Negex* (Chapman et al., 2001) and finally *smoker, current smoker* and *past smoker* categories were identified using lexical and temporal features with the help of an SVM classifier.

4. Methodology

4.1. Extracting Pack-year Information

In the approach we propose, the first step for extracting/quantifying pack-years and classifying instances was to extract portions from text that are relevant to our task. We focus our attention on instance-level extraction, where an instance is a predefined window of characters surrounding a smoking-related term (s-r term). For this purpose, we look at patterns that capture s-r terms such as *smoker*, *tobacco* and *cigarettes* and as there are so many variations to this seemingly small set, we defined five regular expression (regex) patterns in Table 1 to filter these instances.

Table 1. Seed patterns for smoking-related context windows

Patterns for extracting context windows
[Ss][Mm][Oo][Kk]([Ee]|[Ee][Ss]|[Ee][Dd]|[Ii][Nn][Gg]|[Ee][Rr])\b
\b[Tt][Oo][Bb][Aa][Cc][Cc][Oo]\b
\b[Cc][Ii][Gg]([Ss]|[Aa][Rr][Ss]?|[Aa][Rr][Ee][Tt][Tt][Ee][Ss]?)\b
\b[Nn][Ii][Cc][Oo][Tt][Ii][Nn][Ee]\b
\b[Pp][Aa][Cc][Kk][Ss]\b

We selected a sample of 30 documents from the VA document collection and analyzed them to determine the character window size we should utilize for the subsequent modules. We chose a window size of 50 characters on either side of a s-r term. To facilitate a pack-year extraction, we introduce a pattern file for numeric values pre-populated with the most common numeric values and their corresponding patterns. The pre-population is done using patterns for numbers 1 to 100 and major fractions such as 0.5, 0.25, 1.5, etc. This enabled us to map a certain string to an integer or a double value and use them for computing the pack-years figure. This will allow future extension of these patterns if an unseen pattern for a numeric value emerges in the future. A few selected examples of this file is given in Table 2

Table 2. Example patterns in the numeric values pattern file

Value	Corresponding pattern
1.5	[Oo][Nn][Ee](\s{0,2}[Aa][Nn][Dd]\s{0,2}[Aa])?\s{0,2}[Hh][Aa][Ll][Ff]|1\.5|1\s{0,2}1\s?/\s?2
...	...
35	\b35\b|[Tt][Hh][Ii][Rr][Tt][Yy]\s{0,2}[Ff][Ii][Vv][Ee]
...	...
89	\b89\b|[Ee][Ii][Gg][Hh][Tt][Yy]\s{0,2}[Nn][Ii][Nn][Ee]
...	...

Pack-year information usually is expressed as a combination of a frequency term and a duration term. In this context, frequency is how many packs of cigarettes or individual cigarettes a patient smokes per day, per week, per month or per year. Duration is how long the patient has smoked. Although this pattern was visible in many instances we analyzed, the variations in expressing these two together is quite complex. The following examples provides some insight to the examples we encountered (showing the various ways a 30 pack-year history can be mentioned).

3 packs per day for 10 years, *3 ppd for 10 yrs*,
3 pp/day × 10 years, *3 pk/dy × 10 yrs*,
3 p/d ×10YRS, *3 Q.D.*, *30ppy*,
3 packs a day for the last 10 yrs, *30py*, *30pyh*
3 packs/day × ten yrs, *30 pack-year*
3 pcks/day × ten years, *3 packs-per-day × 10 yrs*,
3 packs each day for 10 years, *30 p/year*
three packs a day × 10 years, *three ppd × 10 yrs*

Here ppd denotes packs per day whereas ppy and pyh denotes packs per year and pack-year history respectively. As it can be observed from this small set of examples, the number of ways in which this information is given in EMRs can be quite overwhelming and designing patterns to capture every possible variation is understandably an impossible task. Even though the above example focuses only on a frequency value quantified by *per day* and a duration value of quantified by *years*, the reader should appreciate the fact that the number of other combinations, which we originally left out of this example, makes this a significantly difficult problem to tackle (for example: frequency values quantified by *per week*, *per month*, *per year* and duration values quantified by *weeks*, *months* and *years*). Given the fact that in these examples we have not even provided examples of spelling omissions

and term-boundary issues that is very much encountered in these texts, should be evidence to appreciate the complexity of information extraction task from such sources. For this reason, we focused on generalizing patterns as much as we can and we utilized the fact that most patterns can be splitted into a frequency and a duration pattern in the discussion that follows.

The regex patterns for extracting frequencies was created by disjunctively (OR) appending all the numeric patterns followed by two additional patterns. These additional patterns accounted for packs, cigarettes and cigars as well as time notations. Since the frequency can be expressed in terms of days, weeks, months and years, we designed several patterns to capture this data and appended them to the end of the frequency pattern. Similarly, duration patterns were generated by appending numeric patterns disjunctively as well as adding patterns denoting time towards the end. Since it was observed that most of the times these two portions appear together, one large regex pattern was created by merging these two patterns (frequency + duration). Once a phrase has been extracted using this combined pattern, we ran each of the two individual patterns on the extracted phrase to split it into a frequency and duration phrase. One other reason for doing this was to avoid capturing false positives in terms of the duration pattern alone, since a duration pattern alone could extract some other information piece incorrectly to be a duration phrase according to our definition. For example, consider the following phrase.

he smoked 2 packs a day times 6 years; quit 5 yrs

If we were to extract frequencies and durations separately, *5 yrs* could also be falsely extracted as a duration which in fact, is not. To avoid such situations, we made sure that we extract each frequency and duration together as one unit and later use individual patterns to split the two.

With these two phrases in hand, we make use of our numerical patterns defined earlier to extract the relevant values from the phrases. It is apparent that using a few regex patterns we could determine if the frequency and duration phrases speak of days, weeks, months or years and the pack-year values can be calculated appropriately. We further extracted year-quit patterns as it helps in its own as an important information piece as well as providing valuable insight in classifying each instance into its smoker status label. A similar approach (using appended numeric patterns) was used to produce the regex patterns for extracting year-quit phrases.

4.2. Classification of Smoker Status

The second main task we were faced with is to classify each of these instances within a document into a set of predefined classes. We define four classes, namely, *current smoker*, *past smoker*, *non-smoker* and *unknown* where we strive to follow the rules of the i2b2 challenge described in a previous section. However, we drop the *smoker* class that existed in the i2b2 challenge. In our research, if an instance has a clear mention that a patient has a history of smoking but cannot distinguish with the given amount of information whether the patient is a current or a past smoker, we would like to classify that instance as a current smoker. The class label *unknown* is reserved for instances where there is a mention of a s-r term but it does not specifically relate to the patient under consideration, but is rather a general statement (for example, "*smoking is detremental to your health, if you do smoke, please quit* ").

For the classification task, we make use of Support Vector Machines (SVMs), which is known to be the state-of-the-art in classification methods in the Machine Learning community. An SVM is a construction of a hyperplane or a set of hyperplanes in high dimensional space that achieves highest separation of data points represented by features. Given that the domain under consideration is quite complex and standard rule-based approaches would be difficult to extend, we opted for such a learning approach to facilitate this classification task.

As features for the SVM classification, we defined a combination of word-based-patterns and structure-based-patterns. Again, generalized regular expression patterns that could capture most of the variations in these instances were manually created for this purpose. We came up with 36 word-based-patterns and 70 structure-based-patterns and some of the examples of those features are given in Table 3.

We further make use of the information extracted in the previous section as this can assist immensely in this classification step. Specifically, if there is an year-quit phrase within the instance, we can directly attribute the class label of *past smoker* to that instance whereas if an instance contains a frequency phrase, we are certain that the instance can be classified as either *current smoker* or *past smoker*, which makes it a binary classification problem for such instances. Furthermore, if no year-quit phrase or frequency phrase was extracted from the instance, it could be classified to any of the four classes defined above and thus makes it a multi-class classification problem.

Thus, it is apparent that we could train two sepa-

Table 3. Examples of word-based-pattern features and structure-based-pattern features for smoker status classification

Feature type	Examples
Word-based	[Aa][Gg][Oo], [Pp][Rr][Ee][Ss][Ee][Nn][Tt], [Nn][Ee][Gg][Aa][Tt][Ii][Vv][Ee], [Nn][Oo][Nn]-?\s{0,2}[Ss][Mm][Oo][Kk][Ee][Rr], [Nn][Ee][Vv][Ee][Rr], [Dd][Ee][Nn][Ii][Ee][Ss], [Cc][Uu][Rr][Rr][Ee][Nn][Tt][Ll][Yy], [Cc][Oo][Nn][Tt][Ii][Nn][Uu][Ee][Ss], [Qq][Uu][Ii][Tt], [Dd][Ii][Ss][Cc][Oo][Nn][Tt][Ii][Nn][Uu][Ee][Dd], [Ff][Oo][Rr][Mm][Ee][Rr]([Ll][Yy])?, [Ss][Tt][Oo][Pp][Pp][Ee][Dd], [Ee][Xx]\s*[Ô-9A-Za-z]*\s*[Tt][Oo][Bb][Aa][Cc][Cc][Oo], ..., ..., ..., ..., ..., ...,
Structure-based	s-r term followed by [X]No [ ]Yes (Ex: Do you smoke? [X]No [ ]Yes) s-r term followed by [ ]No [X]Yes (Ex: Do you smoke? [ ]No [X]Yes), s-r term followed by Y/N: N (Ex: Tobacco Y/N: N,) [ ] followed by s-r term (Ex: [ ] Smokes?), [X] followed by s-r term (Ex: [X] Tobacco?), s-r term followed by term denoting negation (Ex: Tobacco-never), ..., ..., ..., ..., ..., ...,

rate SVM models to classify these instances - i.e one binary classifier (SVM-Model-1) to classify instances containing frequency phrases and a multi-class classifier (SVM-Model-2) to classify instances that do not contain any additional information. As such, we can come up with a simple algorithm (Algorithm 1) for classifying the instances.

Algorithm 1 Smoker Status Classification

```
Input: SVM-Model-1, SVM-Model-2,
       Smoking-related context window (x)
if x.contains(year-quit-phrase) then
  x.class = "Past Smoker"
else if !(x.contains(year-quit-phrase))
    and x.contains(frequency-phrase) then
  x.class = SVM-Model-1.predict()
else if !(x.contains(year-quit-phrase))
    and !(x.contains(frequency-phrase)) then
  x.class = SVM-Model-2.predict()
end if
```

5. Evaluation

In this section, we present our experiments in extracting and quantifying pack-years as well as experiments in classification in both document collections. Using the 398 training documents in the i2b2 document set, we were able to extract 272 instances using the five seed patterns for context windows that we defined above. From the 104 test documents, the corresponding number of instances were 84. From the sample documents in the VA document collection, we extracted 1289 training instances as well as 500 test instances.

Table 4. Pack-year extraction results on i2b2 and VA document instances

Dataset	Precision	Recall	F-measure
i2b2	95%(19/20)	95%(19/20)	95.00%
VA	85.34%(99/116)	89.19%(99/111)	87.22%

Only a small number of pack-year mentions were visible in the i2b2 test instances and the precision, recall and f-measure values for this extraction task, along with the corresponding results for the VA instances, are given in Table 4.

As seen from the results, the initial frequency and duration patterns we designed analyzing the training set (with the help of patterns for numeric values) were able to generate and quantify the pack-year values with high precision and recall. An error analysis revealed that the instances which were missed were in fact due to patterns which had not been accounted for before (for example, unusual fractions of packs some patients chose to consume in a day). However, the extendable nature of the pattern files (by adding new numeric patterns, etc) and the supporting architecture allows for incremental improvement until desired accuracy is reached and any missed patterns could be easily added if those patterns are commonly appearing in future documents and are worth including.

We also carried out experiments with the classification task on both document collections. We trained two SVM models (for the two cases described above) using LIBSVM (Chang & Lin, 2001) which is a widely used package for SVMs. We trained the classifiers on

Table 5. Classification accuracies for smoker status on i2b2 and VA document collections

Training set (#instances)	Test set	
	i2b2 (84)	VA (500)
i2b2 (272)	**85.71%**	72.80%
VA (1289)	66.66%	**83.40%**
i2b2 + VA (1561)	77.38%	82.40%

i2b2 data, on VA data and then by combining both training datasets to analyze the difference of accuracies through these combinations. The results are tabulated in Table 5 with the number of instances used depicted within brackets.

The table reveals that the highest accuracy for each category is obtained by using the same category's data for training. This is in line with the common knowledge in similar tasks where classifiers trained on a specific domain tends to do better on unseen data from the same domain. In this case, most of the structured-pattern features that we defined were missing in the i2b2 instances whereas those same features dominated the VA instances. Thus the differences for accuracies can be accounted with a similar reasoning. Also, it is notable that the classifier trained on instances from both these collections does not perform conspicuously worse than classifiers trained on homogeneous data which was a quite interesting observation. This revealed that the classifiers trained on the two diverse document collections performed decently albeit not as well as classifiers trained on individual collections.

6. Concluding Remarks

In this research, we presented a pattern-based approach to extraction, quantification and classification of smoking information in semi-structured medical documents. It was shown that information extraction and classification in these documents is quite a complex task owing to their esoteric nature characterized by clinical sublanguage, idiosyncratic acronyms, term-boundary issues, unforseen templates, etc. We have presented in this paper a methodology that we evaluated on two document collections for extraction, quantification and classification of smoking related instances extracted from medical records. We have shown that simple methods perform significantly well in these complex documents and that the patterns we designed generalize quite nicely to capture most of the desired information from the documents. We further showed that a combination of term-based-patterns and structure-based-patterns can be used a rich feature representation for these instances and the classification results produced are in par with the previous approaches evaluated on less esoteric documents.

References

Aramaki, Eiji, Imai, Takeshi, Miyo, Kengo, and Ohe, Kazuhiko. Patient status classification by using rule based sentence extraction and bm25 knn-based classifier. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

Chang, Chih-Chung and Lin, Chih-Jen. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chapman, Wendy W., Bridewell, Will, Hanbury, Paul, Cooper, Gregory F., and Buchanan, Bruce G. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 5:301–310, 2001.

Clark, Cheryl, Good, Kathleen, Jezierny, Lesley, Macpherson, Melissa, Wilson, Brian, and Chajewska, Urszula. Case report: Identifying smokers with a medical extraction system. *JAMIA*, 15(1):36–39, 2008.

Cohen, Aaron M. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *Journal of the American Medical Informatics Association*, 15(1):32–35, 2008. doi: 10.1197/jamia.M2434.

Ferrucci, David and Lally, Adam. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10:327–348, September 2004. ISSN 1351-3249. doi: 10.1017/S1351324904003523.

Kadunce, Donald P., Burr, Randy, Gress, Richard, Kanner, Richard, Lyon, Joseph L., and Zone, John J. Cigarette smoking: Risk factor for premature facial wrinkling. *Annals of Internal Medicine*, 114(10):840–844, 1991. doi: 10.1059/0003-4819-114-10-840.

Kasiske, Bertram L. and Klinger, Dagmar. Cigarette smoking in renal transplant recipients. *Journal of the American Society of Nephrology*, 11(4):753–759, 2000.

Langhammer, Arnulf, Johnsen, Roar, Holmen, Jostein, Gulsvik, A, and Bjermer, L. Cigarette smoking gives more respiratory symptoms among women than among men the nord-trndelag health study (hunt). *Journal of Epidemiology and Community Health*, 54(12):917–922, 2000. doi: 10.1136/jech.54.12.917.

Pedersen, Ted. Determining smoker status using supervised and unsupervised learning with lexical features. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

Savova, Guergana K, Ogren, Philip V, Duffy, Patrick H, Buntrock, James D, and Chute, Christopher G. Mayo clinic nlp system for patient smoking status identification. *Journal of the American Medical Informatics Association*, 15(1):25–28, 2008. doi: 10.1197/jamia.M2437.

Trevathan, Edwin, Layde, Peter, Webster, Linda A., Adams, Jacob B., Benigno, Benedict B., and Ory, Howard. Cigarette smoking and dysplasia and carcinoma in situ of the uterine cervix. *JAMA: The Journal of the American Medical Association*, 250(4):499–502, 1983. doi: 10.1001/jama.1983.03340040039028.

Uzuner, zlem, Goldstein, Ira, Luo, Yuan, and Kohane, Isaac. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008. doi: 10.1197/jamia.M2408.

Valdes, AM, Andrew, T, Gardner, JP, Kimura, M, Oelsner, E, Cherkas, LF, Aviv, A, and Spector, TD. Obesity, cigarette smoking, and telomere length in women. *The Lancet*, 366(9486):662 – 664, 2005. ISSN 0140-6736. doi: DOI:10.1016/S0140-6736(05)66630-5.

Zeng, Qing, Goryachev, Sergey, Weiss, Scott, Sordo, Margarita, Murphy, Shawn, and Lazarus, Ross. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6(1):30, 2006. ISSN 1472-6947. doi: 10.1186/1472-6947-6-30.

Zhou, Xiaohua, Han, Hyoil, Chankai, Isaac, Prestrud, Ann, and Brooks, Ari D. Converting semi-structured clinical medical records into information and knowledge. In *ICDE Workshops*, pp. 1162, 2005.